* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1] The speech recognition approach which creates the model of an unspecified speaker and is characterized by recognizing using this in the speech recognition approach using a continuous-distribution mold hidden Markov model by creating a model for every speaker, mixing output distribution of all speakers for every condition of the same voice, and considering as mixed distribution in case a model is created with two or more speakers' voice.
[Claim 2] The speech recognition approach of claim 1 characterized by replacing with said continuous-distribution mold hidden Markov model, and two or more phonemic models sharing a condition in a context dependence phonemic model.
[Claim 3] The speech recognition approach of claims 1 or 2 characterized by using a movement vector place smoothing method and creating a model in case the model for every speaker is created with two or more of said speakers' voice.
[Claim 4] The speech recognition approach of either [which carries out the relearning of the mixed multiplier using input voice, and is characterized by using the model for recognition under constraint of making the mixed multiplier applied to the output distribution acquired from the same speaker in said obtained unspecified speaker phonemic model into the same value] claim 1 thru/or the either of 3.
[Claim 5] The speech recognition approach of claim 4 which reduces the number of mixing and is characterized by to use for recognition the model obtained by carrying out reallocation of the weight so that the sum of the weight of mixed output distribution may be set to 1 after that by removing the mixed element with which the mixed multiplier of a model became below a threshold from a model as a result of carrying out the relearning, when determining the threshold of said mixed multiplier.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Industrial Application] Especially this invention relates to the speech recognition approach like the speaker adaptation speech recognition which made the early model the speech recognition and the unspecified speaker phonemic model for an unspecified speaker about the speech recognition approach.
[0002]
[Description of the Prior Art] At the former, in order to recognize the voice of an unspecified speaker, the phonemic model of an unspecified speaker was created by learning a phonemic model, without distinguishing a speaker. This approach was what does not use constraint that one utterance is a thing from the same speaker.
[0003] Thus, by the approach of using for phoneme study, without distinguishing two or more speakers' voice, distribution of each phoneme spreads, and it laps with other phonemes greatly, and is connected with incorrect recognition. For example, Speaker's A /a/may be the same as Speaker's B /o/, or distribution may lap greatly. Moreover, in order to improve the recognition engine performance, when the number of mixing of a phonemic model was made to increase, there was a fault that computational complexity increased.
[0004] So, the main purpose of this invention is offering the speech recognition approach which can acquire a high recognition rate by little data, and can reduce computational complexity.
[0005]
[Means for Solving the Problem] This invention is the speech recognition approach which used the continuous-distribution mold hidden Markov model, in case it creates a model with two or more speakers' voice, by creating a model for every speaker, mixing output distribution of all speakers for every condition of the same voice, and considering as mixed distribution, creates the model of an unspecified speaker and recognizes using this.
[0006] In a context dependence phonemic model, two or more phonemic models share a condition instead of a continuous-distribution mold hidden Markov model more preferably.
[0007] Furthermore, in case the model for every speaker is created with two or more more desirable speakers' voice, a movement vector place smoothing method is used and a model is created.
[0008] Furthermore, in the unspecified speaker phonemic model obtained more preferably, under constraint of making into the same value the mixed multiplier concerning the output distribution acquired from the same speaker, the relearning of the mixed multiplier is carried out using input voice, and the model is used for recognition.
[0009] When determining the threshold of a mixed multiplier, as a result of carrying out the relearning further more preferably, by removing the mixed element with which the mixed multiplier of a model became below a threshold from a model, the number of mixing is reduced and the model obtained by carrying out reallocation of the weight so that the sum of the weight of mixed output distribution may be set to 1 after that is used for recognition.
[0010]
[Function] By the speech recognition approach concerning this invention creating a model for

every speaker, mixing output distribution of all speakers for every condition of the same phoneme, and considering as mixed distribution By being able to prevent confusion of the phoneme between speakers to voice model creation time using constraint of speaker coordination, and controlling only probability weight, speaker adaptation can be performed with very little input voice, and computational complexity at the time of recognition can be reduced.
[0011]
[Example] Drawing 1 is the outline block diagram of one example of this invention. The speech recognition method of this invention is used for example, for an automatic translation telephone, and as shown in drawing 1 , it consists of amplifier 1, a low pass filter 2, A/D converter 3, and a processor 4. Amplifier 1 amplifies the inputted sound signal and a low pass filter 2 removes a noise from the amplified sound signal repeatedly. A/D converter 3 changes a sound signal into a 16-bit digital signal with a 12kHz sampling signal. A processor 4 contains a computer 5, a magnetic disk 6, terminals 7, and a printer 8. A computer 5 performs speech recognition based on the process memorized by the magnetic disk 6 based on the audio digital signal inputted from A/D converter 3.
[0012] Drawing 2 is a flow chart for explaining actuation of one example of this invention. Next, actuation of one example of this invention is explained with reference to drawing 1 and drawing 2 . By the approach by this invention, after creating the unspecified speaker phonemic model by speaker mixing, speaker adaptation by speaker weight study is performed, and speaker pruning is performed after that.
[0013] First, in creation of the unspecified speaker phonemic model by speaker mixing, an unspecified speaker phonemic model is created by using the phonemic model created for every speaker as a mixed component of an unspecified speaker phonemic model. The first stage [as / a lot of data of one speaker memorized by the magnetic disk 6 in the step (it is called SP for short in illustration) SP 1 shown in drawing 2 to whose output distribution is single Gaussian distribution] HMnet Successive StateSplitting It generates using an algorithm (SSS). HMnet and an SSS algorithm -- Takami and Sagayama -- "-- it is related with a phoneme context and time amount -- serially -- a part for a condition -- it depends comparatively -- it can hide and automatic generation" of the Markov network, the Institute of Electronics, Information and Communication Engineers voice study group data, and SPs 91-88 (December, 1991) can be used. This model can be used as speech recognition of a specified speaker.
[0014] Next, parameter study is performed. In the above-mentioned step SP 1, after an SSS algorithm determines the structure of HMnet, as shown in a step SP 2, it asks for the parameter of HMnet for every speaker from comparatively little two or more speakers' voice data memorized by the magnetic disk 6. A movement vector place smoothing method (Vector Field Smoothig:VFS) is used as an approach of a parameter. About this VFS, Okura, Sugiyama, Sagayama, and "movement vector place smoothing speaker adaptation method using mixed continuous distribution HMM" Institute of Electronics, Information and Communication Engineers voice study group data, SP 92-16, and 23rd page - the 28th page (June, 1992) can be used. Thus, HMnet which was adapted for two or more speakers is generated by preparing two or more speakers' voice data, and performing study of the parameter of HMnet using the VFS method for every speaker.
[0015] Next, speaker mixing-ization is performed. The mixed consecutive output distribution HMnet is created by expressing the condition that HMnet for two or more speakers corresponds to a step SP 3 so that it may be shown, as one mixed output distribution. Speaker mixing-ization is performed in HMnet with the same structure by summarizing to one the output distribution which the condition of being in the same location among structure has, and expressing as mixed consecutive output distribution. a branching probability -- same probability -- or -- Baum-Welch The relearning only of the branching probability is carried out and it is determined by the algorithm.
[0016] Drawing 3 is the conceptual diagram of speaker mixing, speaker weight study, and speaker pruning, and especially drawing 3 (a) shows the concept of above-mentioned speaker mixing, and shows the output distribution of Speakers A, B, and C in Condition i and Condition j.
[0017] next, the speaker adaptation by speaker weight study is attached [it is alike and] and

explained. The mixed continuous distribution HMnet acquired by above-mentioned explanation are used as the base, and the technique of performing speaker adaptation with a small number of input voice is explained. In the mixed consecutive output distribution HMnet created by the speaker mixing SSS, the origin whether the mixed component which constitutes each mixed output distribution is generated from which speaker's data is known. Therefore, the branching probability to each mixed component can be understood to be a weighting factor to each speaker. For this reason, it is possible to treat the branching probability concerning the mixed component originating in the same speaker, i.e., a speaker weighting factor, as an "epilogue." Using the property of this speaker mixing SSS, as shown in a step SP 4, speaker adaptation by speaker weight study is performed. First, the average, distribution, and transition probability of output distribution are not updated, but updates the weight to the mixed element accepted by "the epilogue, i.e., the same speaker," among speakers only in the weighting factor using a Baum-Welch algorithm under the constraint of making it the same. Connection study is used for this study. The concept of above-mentioned speaker weight study is shown in drawing 3 (b). Thus, speech recognition is performed using the model which was adapted for an input speaker's voice.

[0018] Next, speaker pruning is explained. When it becomes below the probability for the weighting factor to have been beforehand set up by speaker weight study among mixed output distribution of HMnet, in a step SP 5, the weighting factor is transposed to 0. Then, reallocation of the weight is carried out so that the sum of the weight of mixed output distribution may be set to 1. The principle is shown in drawing 3 (c). As shown in drawing 3 (c), a model is simplified by deleting all the small mixing components of the speaker weight originating in the same speaker. Thus, speech recognition is performed using the phonemic model to which size was reduced.

[0019] Drawing 4 is the speaker adaptation by the speaker weight study by one example of this invention, speaker weight study + speaker pruning, and drawing showing the result of a recognition experiment of three kinds of Japanese clause recognition of all weight study. When the result of speaker adaptation is seen, as for any speaker, it understands that improvement in a recognition rate is obtained with very few samples of 1 - 5 word extent. By the approach of learning all the weight of the conventional approach independently, since there are many parameters for study, if there are few study words, a recognition rate will fall conversely.

[0020] Change of the number of mixing at the time of performing speaker pruning is shown in Table 1. Thus, although the number of output distribution decreases to about 1 / two to 1/12 by each speaker, especially the decline in a recognition rate is not seen.

[0021]

[Table 1]

発話者	# 学習単語					
	0	1	5	10	30	60
MMS	12	4	3	5	6	5
MMY	12	2	1	1	1	1
MSH	12	2	2	1	1	1

[0022]

[Effect of the Invention] As mentioned above, according to this invention, in the speech recognition approach using a continuous-distribution mold hidden Markov model, in case a model is created with two or more speakers' voice, a high recognition rate can be acquired by little data

by creating the model of an unspecified speaker and recognizing using this by creating a model for every speaker, mixing output distribution of all speakers for every condition of the same phoneme, and considering as mixed distribution.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1] It is the outline block diagram of one example of this invention.
[Drawing 2] It is a flow chart for explaining actuation of one example of this invention.
[Drawing 3] It is the conceptual diagram of speaker mixing by this invention, speaker weight study, and speaker pruning.
[Drawing 4] It is drawing showing the clause recognition experimental result after the speaker adaptation by this invention.
[Description of Notations]
1 Amplifier
2 Low Pass Filter
3 A/D Converter
4 Processor
5 Computer
6 Magnetic Disk
7 Terminals
8 Printer

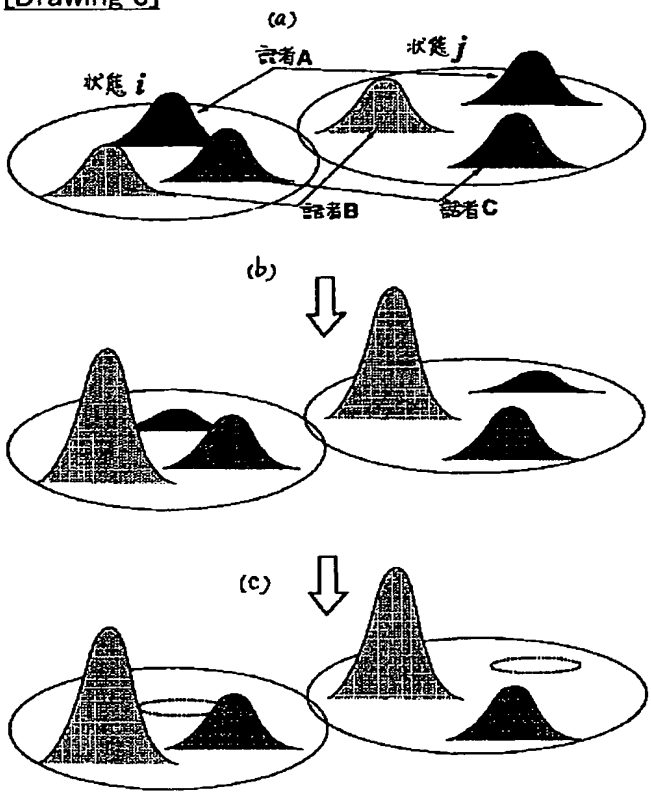
[Translation done.]

* NOTICES *

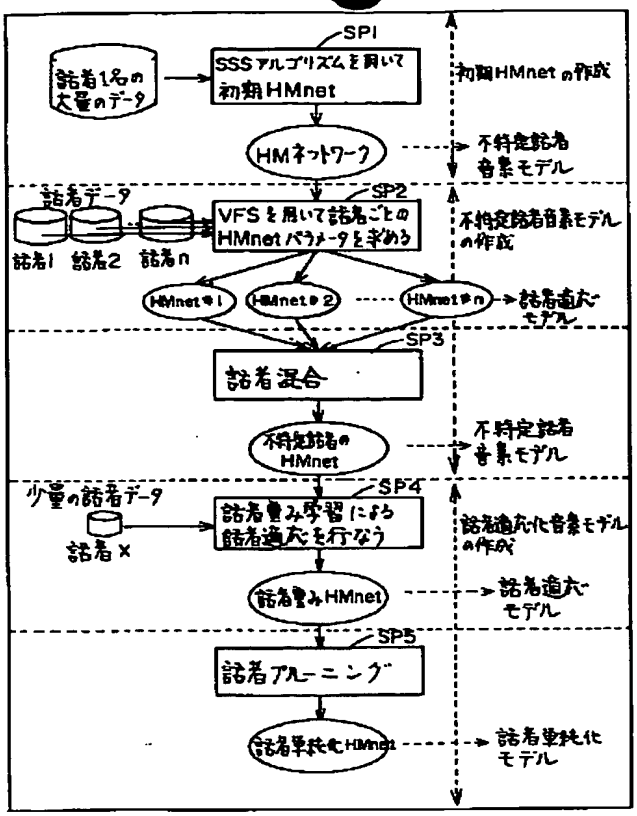
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

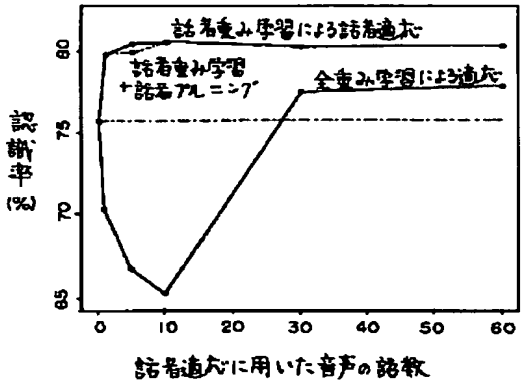
[Drawing 3]



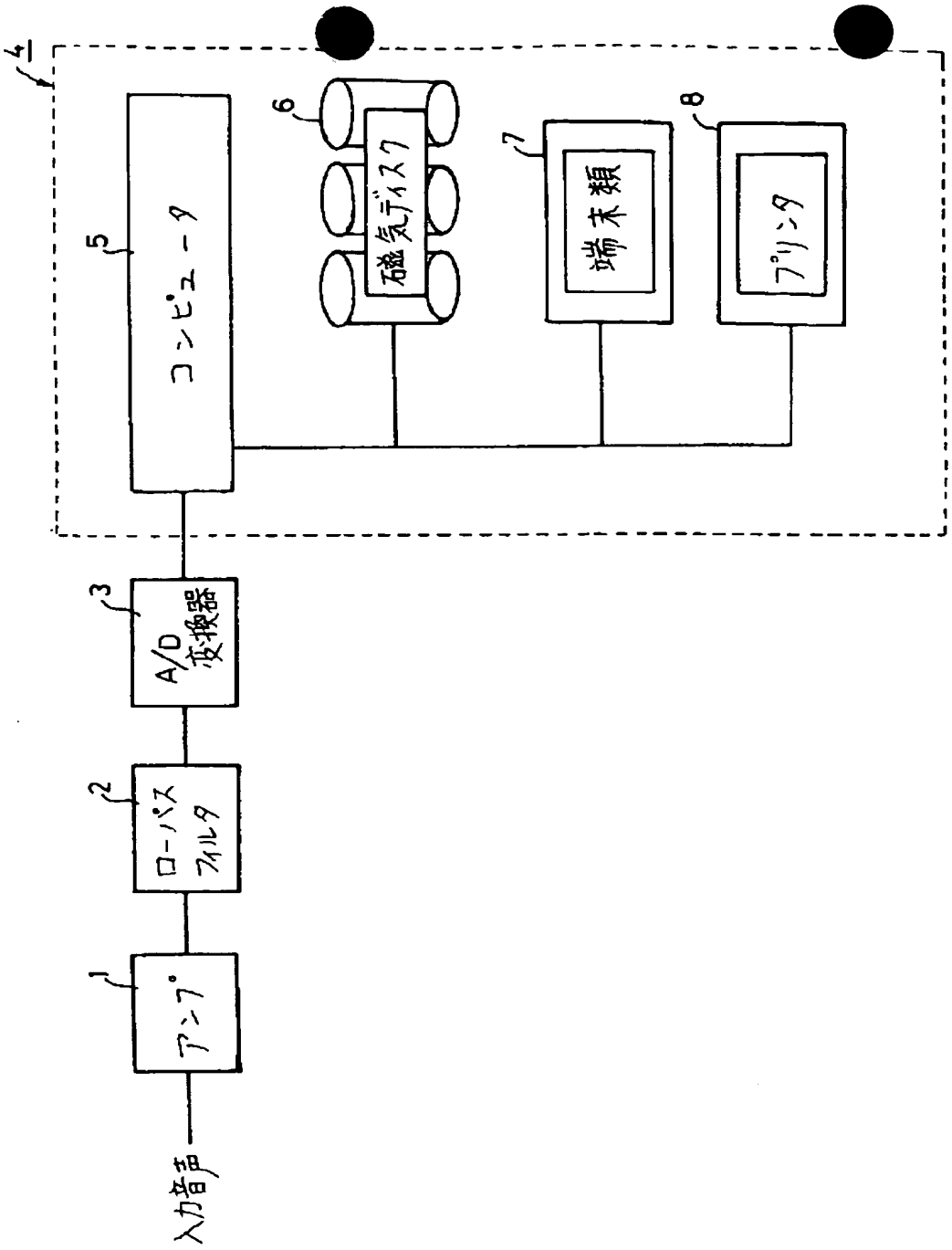
[Drawing 2]



[Drawing 4]



[Drawing 1]



[Translation done.]

全項目

(19)【発行国】日本国特許庁(JP)
(12)【公報種別】特許公報(B2)
(11)【公告番号】特公平7－69711
(24)(44)【公告日】平成7年(1995)7月31日
(54)【発明の名称】音声認識方法
(51)【国際特許分類第6版】

G10L 3/00 535
521 C
F

【請求項の数】5
【全頁数】5
(21)【出願番号】特願平5－48271
(22)【出願日】平成5年(1993)3月9日
(65)【公開番号】特開平6－259089
(43)【公開日】平成6年(1994)9月16日
【新規性喪失の例外の表示】特許法第30条第1項適用申請有り 1992年9月10日、社団法人電子情報通信学会発行の「電子情報通信学会技術研究報告Vol. 92No. 207」に発表
(71)【出願人】
【識別番号】000127684
【氏名又は名称】株式会社エイ・ティ・アール自動翻訳電話研究所
【住所又は居所】京都府相楽郡精華町大字乾谷小字三平谷5番地
(72)【発明者】
【氏名】嵯峨山 茂樹
【住所又は居所】京都府相楽郡精華町大字乾谷小字三平谷5番地 株式会社エイ・ティ・アール自動翻訳電話研究所内
(74)【代理人】
【弁理士】
【氏名又は名称】深見 久郎 (外2名)
【審査官】橋本 武

【特許請求の範囲】
【請求項1】 連続分布型隠れマルコフモデルを用いた音声認識方法において、複数の話者の音声によりモデルを作成する際に、話者ごとにモデルを作成し、同一音声の各状態ごとに全話者の出力分布を混合して混合分布とすることにより、不特定話者のモデルを作成し、これを用いて認識を行なうことを特徴とする、音声認識方法。
【請求項2】 前記連続分布型隠れマルコフモデルに代えてコンテキスト依存音素モデルにおいて複数の音素モデルが状態を共有することを特徴とする、請求項1の音声認識方法。
【請求項3】 前記複数の話者の音声により話者ごとのモデルを作成する際に、移動ベクトル場平滑化方式を用いてモデルを作成することを特徴とする、請求項1または2の音声認識方法。
【請求項4】 前記得られた不特定話者音素モデルにおいて、同一話者から得られた出力分布にかかる混合係数を同一の値にするという制約の下で、入力音声を用いて混合係数を再学習し、そのモデルを認識に用いることを特徴とする、請求項1ないし3のいずれかの音声認識方法。
【請求項5】 前記混合係数のしきい値を決定する際に、再学習した結果、モデルの混合係数がしきい値以下になった混合要素をモデルから除くことにより混合数を減らし、その後混合出力分布の重みの和が1となるように重みを再配分することにより得られたモデルを認識に用いることを特徴とする、請求項4の音声認識方法。

【発明の詳細な説明】
【0001】
【産業上の利用分野】この発明は音声認識方法に関し、特に、不特定話者を対象とした音声認識および不特定話者音素モデルを初期モデルとした話者適応音声認識のような音声認識方法に関する。
【0002】
【従来の技術および発明が解決しようとする課題】従来では、不特定話者の音声を認識するために、話者を区別することなく音素モデルを学習することで、不特定話者の音素モデルを作成していた。この方法は、1つの発声は同一の話者からのものである、という制約を利用しないものであった。
【0003】このように、複数の話者の音声を区別することなく音素学習に用いる方法では、各音素の分布が広がって他の音素と大きく重なり、誤認識に結び付く。たとえば、話者Aの／a／は、話者Bの／o／と同じであったりあるいは分布が大きく重なることがあり得る。また、認識性能を上げるために、音素モデルの混合数を増加させると、計算量が増大するという欠点があった。
【0004】それゆえに、この発明の主たる目的は、少ないデータで高い認識率を得ることができかつ計算量を削減できるような音声認識方法を提供することである。
【0005】
【課題を解決するための手段】この発明は連続分布型隠れマルコフモデルを用いた音声認識方法であって、複数の話者の音声によりモデルを作成する際に、話者ごとにモデルを作成し、同一音声の各状態ごとに全話者の出力分布を混合して混合分布とすることにより、不特定話者のモデルを作成し、これを用いて認識を行なう。
【0006】より好ましくは、連続分布型隠れマルコフモデルの代わりに、コンテキスト依存音素モデルにおいて複数の音素モデルが状態を共有する。
【0007】さらに、より好ましくは複数の話者の音声により話者ごとのモデルを作成する際に、移動ベクトル場平滑化方式を用いてモデルを作成する。
【0008】さらに、より好ましくは得られた不特定話者音素モデルにおいて、同一話者から得られた出力分布にかかる混合係数を同一の値にするという制約の下で、入力音声を用いて混合係数を再学習し、そのモデルを認識に用いる。
【0009】さらにより好ましくは、混合係数のしきい値を決定する際に、再学習した結果、モデルの混合係数がしきい値以下になった混合要素をモデルから除くことにより混合数を減らし、その後混合出力分布の重みの和が1となるように重みを再配分することにより得られたモデルを認識に用いる。
【0010】
【作用】この発明に係る音声認識方法は、話者ごとにモデルを作成し、同一音素の各状態ごとに全話者の出力分布を混合して、混合分布とすることにより、話者一貫性の制約を音声モデル作成時に利用して、話者間での音素の混同を防ぐことができ、確率重みのみを制御することにより、非常に少ない入力音声で話者適応を行なうことができ、認識時の計算量の削減を行なえる。
【0011】
【実施例】図1はこの発明の一実施例の概略ブロック図である。この発明の音声認識法はたとえば自動翻訳電話に用いられるものであり、図1に示すように、アンプ1とローパスフィルタ2とA／D変換器3と処理装置4とから構成されている。アンプ1は入力された音声信号を増幅し、ローパスフィルタ2は増幅された音声信号から繰り返し雑音を除去する。A／D変換器3は音声信号をたとえば12kHzのサンプリング信号により、16ビットのデジタル信号に変換する。処理装置4はコンピュータ5と磁気ディスク6と端末類7とプリンタ8とを含む。コンピュータ5はA／D変換器3から入力された音声のデジタル信号に基づいて、磁気ディスク6に記憶されているプロセスに基づいて、音声認識を行なう。
【0012】図2はこの発明の一実施例の動作を説明するためのフローチャートである。次に、図1および図2を参照して、この発明の一実施例の動作について説明する。この発明による方法では、話者混合による不特定話者音素モデルの作成を行なった後、話者重み学習による話者適応を行ない、その後話者プルーニングを行なう。
【0013】まず、話者混合による不特定話者音素モデルの作成では、話者ごとに作成した音素モデルを不特定話者音素モデルの混合成分として用いることにより、不特定話者音素モデルが作成される。図2に示すステップ（図示ではSPと略称する）SP1において、磁気ディスク6に記憶されてい

る話者1名の大量のデータから、出力分布が単一ガウス分布であるような初期HMnetをSuccessive StateSplitting（SSS）アルゴリズムを用いて生成する。HMnetおよびSSSアルゴリズムについては、鷹見，嵯峨山，「音素コンテキストと時間に関する逐次状態分割による隠れマルコフ網の自動生成」，電子情報通信学会音声研究会資料，SP91～88（1991年12月）を用いることができる。このモデルは特定話者の音声認識として用いることができる。

【0014】次に、パラメータ学習を行なう。上述のステップSP1において、SSSアルゴリズムによりHMnetの構造を決定した後に、ステップSP2に示すように、磁気ディスク6に記憶されている比較的少量の複数話者の音声データより話者ごとのHMnetのパラメータを求める。パラメータの学習法として移動ベクトル場平滑化方式（Vector Field Smoothig ：VFS）を用いる。このVFSについては大倉，杉山，嵯峨山，「混合連続分布HMMを用いた移動ベクトル場平滑化話者適応方式」，電子情報通信学会音声研究会資料，SP92－16，第23頁～第28頁（1992年6月）を用いることができる。このようにして、複数話者の音声データを用意し、話者ごとにHMnetのパラメータの学習をVFS法を用いて行なうことにより、複数話者に適応したHMnetが生成される。

【0015】次に、話者混合化を行なう。ステップSP3に示すように、複数話者分のHMnetの対応する状態を1つの混合出力分布として表現することにより、混合連続出力分布HMnetが作成される。話者混合化は、同一の構造を持つHMnetにおいて、構造中同一の位置にある状態が持つ出力分布を1つにまとめ混合連続出力分布として現わすことにより行なわれる。分岐確率は等確率または Baum-Welch アルゴリズムによって分岐確率のみ再学習して決定される。

【0016】図3は話者混合，話者重み学習および話者プルーニングの概念図であり、特に、図3（a）は上述の話者混合の概念を示したものであり、状態iと状態jとにおける話者A，B，Cの出力分布を示す。

【0017】次に、話者重み学習による話者適応についてについて説明する。上述の説明によって得られた混合連続分布HMnetをベースにして、少数の入力音声により話者適応を行なう手法について説明する。話者混合SSSで作成された混合連続出力分布HMnetでは、各混合出力分布を構成する混合成分はどの話者のデータから生成されたものであるかという由来がわかっている。したがって、各混合成分への分岐確率は各話者への重み係数と理解できる。このため、同一話者に由来する混合成分にかかる分岐確率、つまり話者重み係数を「結び」として扱うことが可能である。この話者混合SSSの性質を利用して、ステップSP4に示すように、話者重み学習による話者適応を行なう。まず、出力分布の平均値・分散・遷移確率は更新せず、重み係数のみを話者間で「結び」、つまり同一話者から認められた混合要素に対する重みを同一にするという拘束条件の下でBaum-Welchアルゴリズムを用いて更新する。この学習には連結学習を用いる。上述の話者重み学習の概念を図3（b）に示す。このようにして、入力話者の音声に適応されたモデルを用いて音声認識が行なわれる。

【0018】次に、話者プルーニングについて説明する。HMnetの混合出力分布のうち、話者重み学習により重み係数が予め設定された確率以下になった場合、ステップSP5においてその重み係数を0に置換える。その後、混合出力分布の重みの和が1となるように重みを再配分する。その原理を図3（c）に示す。図3（c）に示すように、同一話者に由来する話者重みの小さい混合成分をすべて削除することによりモデルの単純化を行なう。このようにしてサイズが縮小された音素モデルを用いて音声認識が行なわれる。

【0019】図4はこの発明の一実施例による話者重み学習による話者適応，話者重み学習＋話者プルーニング，全重み学習の3種類の日本語文節認識の認識実験の結果を示す図である。話者適応の結果を見ると、いずれの話者も1～5単語程度の非常に少ないサンプルで認識率の向上が得られることがわかる。従来の方法のすべての重みを独立に学習する方法では、学習対象パラメータが多いために学習単語が少ないと逆に認識率が低下する。

【0020】話者プルーニングを行なった場合の混合数の変化を表1に示す。このように各話者で出力分布数が1／2～1／12程度に減少するが、特に認識率の低下は見られない。

【0021】

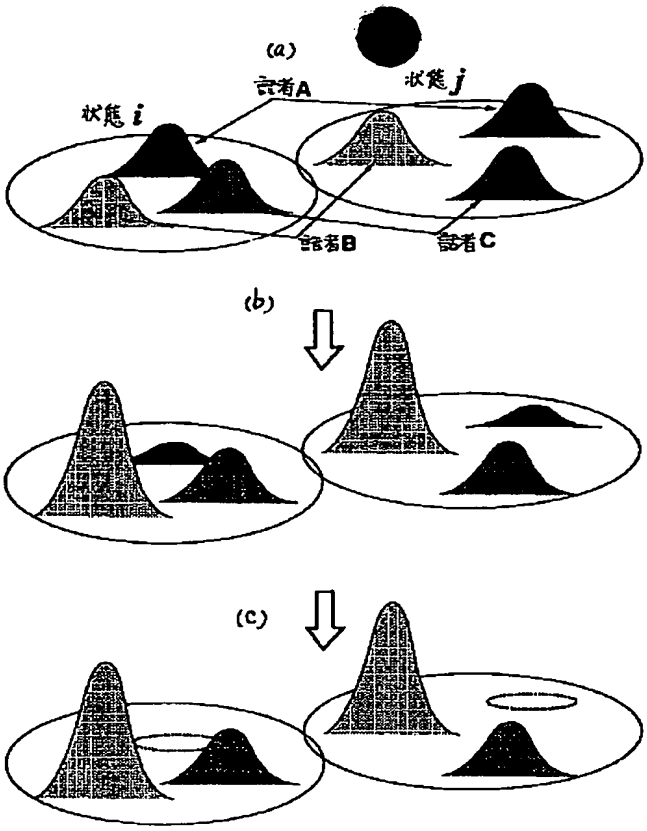
【表1】

発話者	# 学習単語					
	0	1	5	10	30	60
MMS	12	4	3	5	6	5
MMY	12	2	1	1	1	1
MSH	12	2	2	1	1	1

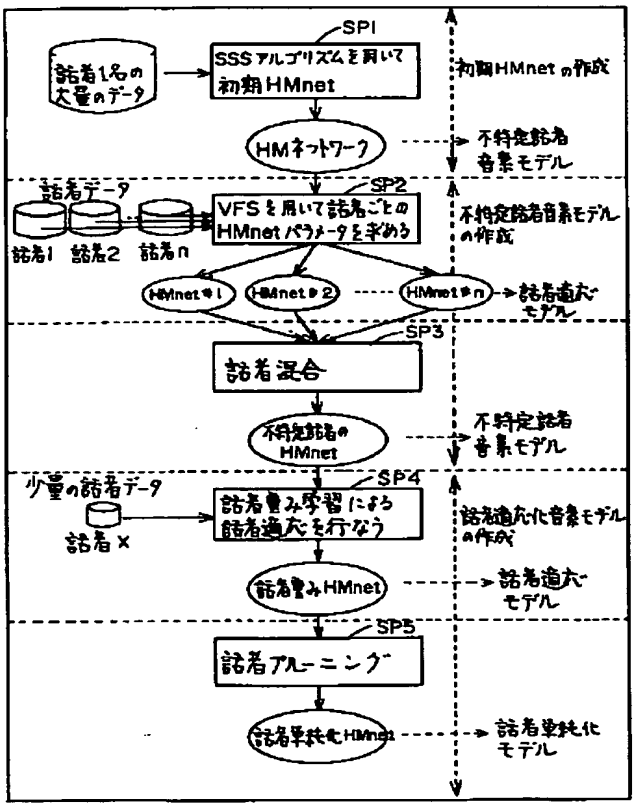
【0022】
【発明の効果】以上のように、この発明によれば、連続分布型隠れマルコフモデルを用いた音声認識方法において、複数の話者の音声によりモデルを作成する際に、話者ごとにモデルを作成し、同一音素の各状態ごとに全話者の出力分布を混合して混合分布とすることにより、不特定話者のモデルを作成し、これを用いて認識を行なうことにより、少ないデータで高い認識率を得ることができる。

【図面の簡単な説明】
【図1】この発明の一実施例の概略ブロック図である。
【図2】この発明の一実施例の動作を説明するためのフローチャートである。
【図3】この発明による話者混合，話者重み学習，話者プルーニングの概念図である。
【図4】この発明による話者適応後の文節認識実験結果を示す図である。
【符号の説明】
1 アンプ
2 ローパスフィルタ
3 A／D変換器
4 処理装置
5 コンピュータ
6 磁気ディスク
7 端末類
8 プリンタ

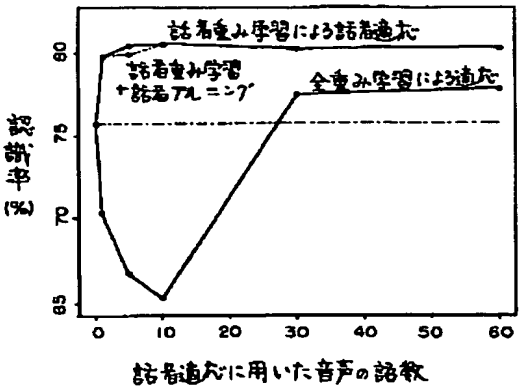
【図3】



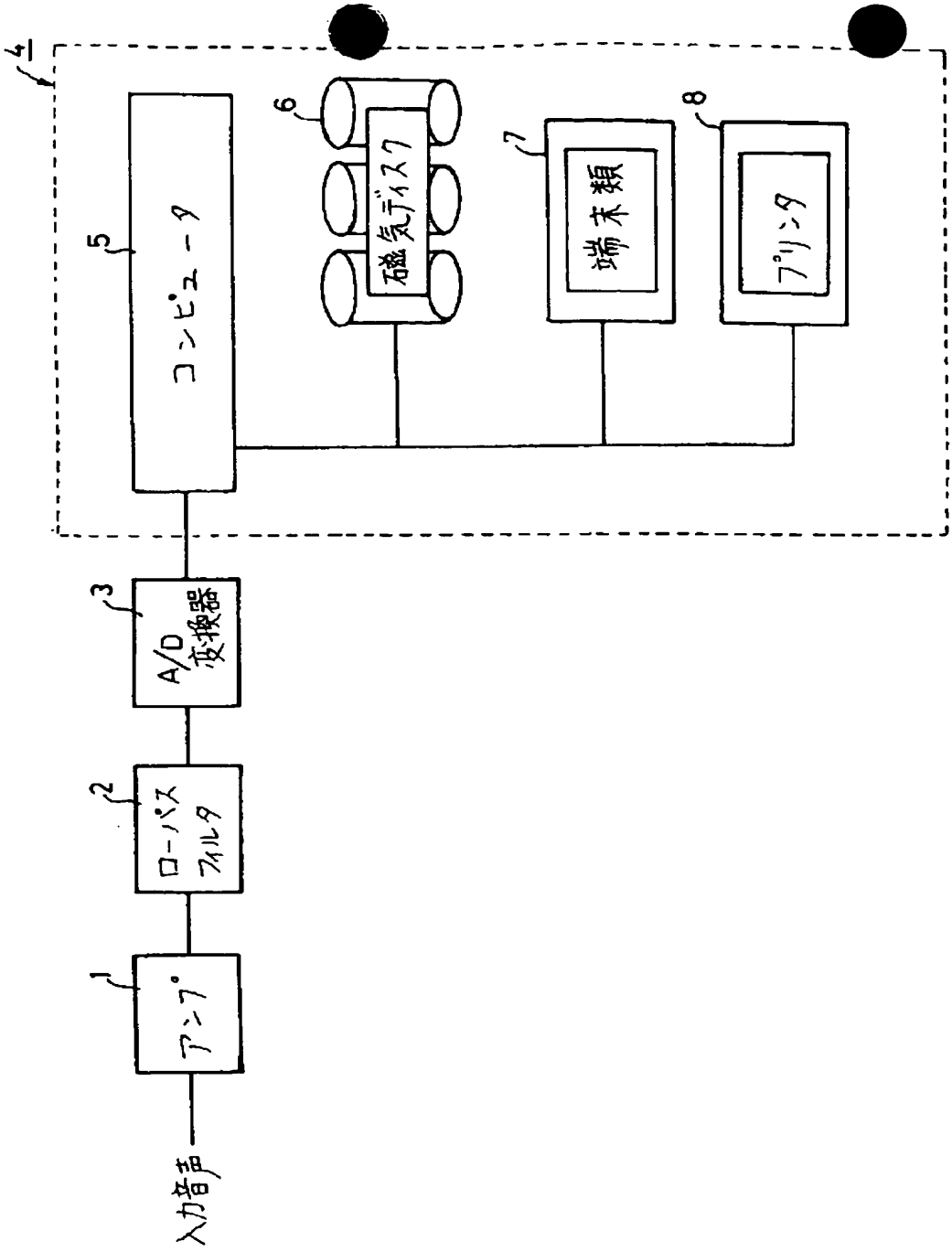
【図2】



【図4】



【図1】


4
5
コンピュータ
6
磁気ディスク
7
端末類
8
プリンタ
3
A/D
変換器
2
ローパス
フィルタ
1
アンプ
入力音声